САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Математико-механический факультет

Кафедра Системного Программирования

Соса Екатерина Андреевна

Исследование задачи поиска генов в метагеномных сборках de novo

Курсовая работа

Научный руководитель:

Минкин И. В.

Санкт-Петербург
2013

Оглавление

Введение

Вычислительная биология или биоинформатика – наука, решающая задачи на стыке биологии, информатики и математики. Одна из больших её областей – исследование геномов организмов. Эта область включает в себя и решение задач, связанных с биотехнологическими процессами "оцифровки" геномной информации об организмах, и анализ полученных последовательностей. Решением биотехнологических задач занимаются, скорее, биологи и химики, а помогают им в этом инженеры и физики, тогда как анализировать результаты их работы необходимо математикам и программистам, интересующимся биологией.

Наиболее частым и реальным применением результатов исследований в области биоинформатики принято считать персональную медицину и фармакологию, а борятся учёные за дешевизну (в денежном смысле) и скорость работы (сейчас время работы многих программ насчитывает часы, сутки и даже месяцы, несмотря на возможности современных компьютеров).

Геном – уникальная информация, характеризующая организм и извлекаемая из ДНК. Надо понимать, что геном сам по себе не несёт информации о болезнях и других особенностях, а значит мало расшифровать сам геном, нужно уметь анализировать представленные в нём данные и правильно их интерпретировать. Долгое время было принято считать, что вся нужная для этого информация хранится в небольших участках, в генах, хотя они составляют единичный процент всего генома. Позже стало понятно, что изучать нужно весь геном, а не отдельные его части, но ценность поиска и анализа генов от этого не упала, ведь модификации именно в них обусловливают характерные черты отдельной особи.

На самом деле, анализ и сборка геномных последовательностей – это лишь часть задач, рассматриваемых в биоинформатике, но и этот набор достаточно широк. В данной работе рассмотрена задача поиска генов в метагеномных сборках de novo. Несмотря на то, что используемые термины будут определены более формально, стоит пояснить некотрые из них для постановки задачи.

Геном – строка в алфавите {A, C, G, T}, характеризующая конкретный организм.

Сборка – набор подстрок генома, подстроки небольшой длины (длины, которая на

порядки меньше общей длины генома). При этом объединение элементов сборки не обязательно покрывает весь геном, а также элементы могут накладываться друг на друга и содержать ошибки.

Метагеномная сборка – смесь сборок нескольких геномов, то есть набор небольших подстрок нескольких геномов.

Ген – участок (подстрока) генома, кодирущий белок, поэтому и отличающийся от остальных участков определёнными особенностями. Эти особенности не определяют участок однозначно и могут изменяться от класса организмов к другому классу, а иногда и от организма к организму.

de novo сборка означает, что полные геномные последовательности организмов, находящихся в метагеномной сборке, ещё не известны.

Исходя из определений, можно выделить несколько особенностей, которые и становятся основой исследования поставленной задачи. Во-первых, гены обладают особенностями, но эти особенности не задают нам конкретной строки, которую необходимо найти, поэтому задача их поиска не сводится к поиску подстроки в строке, причём ни к точному, ни к неточному. Как следствие, корректнее назвать эту задачу задачей предсказывания генов, так как в качестве решения предоставляется набор строк, но проверить, является ли каждая из этих строк реальным геном, возможно только по результатам биологического эксперимента. Во-вторых, если гены обладают особенностями, которые можно определить, зная, для какого организма их необходимо найти, то в случае метагномной сборки это не поможет, так как там присутствуют части геномов различных организмов. В-третьих, задачи анализа метагеномной информации упрощаются, если биологи когда-то уже получили полную последовательность для одного или нескольких организмов, присутствующих в сборке: эту последовательность можно использовать при исследовании. Однако в текущей формулировке подразумевается, что никаких данных, известных заранее, нет, либо о них не известно.

1. Постановка задачи

1.1. Основные определения

Прежде чем формулировать задачу, необходимо определить основные понятия и дать небольшое введение в предметную область. Расшифровки билогических понятий даны в двух интерпретациях: первая – биологическое определение, взятое из указанных источников, а вторая – определение, близкое к математическому и непосредственно используемое в данной работе.

Геном (последовательность ДНК) – 1) это длинный неразветвлённый полимер, состоящий всего из четырех субъединиц-дезоксирибонуклеотидов, азотистые основания которых представлены аденином (A), цитозином (C), гуанином (G) и тимином (T); нуклеотиды связаны между собой ковалентными фосфодиэфирными связями, соединяющими 5'-атом углерода одного остатка с 3'-атомом углерода следующего остатка[16];
2) строка в алфавите {A, C, G, T}, однозначно определяющая организм (или штамм), к которому она относится.

Стоит сказать, что геномы, хоть они и уникальны, у близких организмов похожи. Насколько между собой похожи геномы, показывают эволюционные деревья – деревья, характеризующие ход эволюции.

Задача полной расшифровки генома, то есть представления полной геномной последовательности, до сих пор под силу только биологам, так как в последней части её решения необходим биологический эксперимент. Если рассматривать процесс расшифровки генома полностью, то он состоит из нескольких последовательных частей: секвенирование, ассемблирование и финишинг. На этапе секвенирования при помощи различных биотехнологических процессов из биологического материала получаются риды – последовательности букв одинаковой длины (обычно 100) из алфавита {A, C, G, T}, далее при помощи специальных программ-ассемблеров эти риды склеиваются в более длинные строки, которые называются контигами, однако они всё равно не покрывают весь геном и не несут информацию о том, где именно в геноме они расположены. Результат работы стадии ассемблирования – множество полученных контигов – называется сборкой. Интересно, что для дальнейшего анализа генома можно использовать результат

ассемблирования.

В зависимости от биологического материала, из которого получаются данные, сборки бывают single-cell, multi-cell и метагеномными.

Single-cell сборка – сборка, для получения которой была взята одна или несколько (до четырёх) клеток колонии.

Multi-cell сборка или просто сборка подразумевает, что в качестве биологического материала были взяты тысячи или даже десятки тысяч клеток одного штамма.

Метагеномная сборка (иногда говорят "мультигеномная") – сборка, для которой взяли не выделенные клетки одного штамма[25], а среду обитания целевой колонии, в которой были как её представители, так и соседствующих. В такой сборке присутствуют геномы нескольких штаммов, но обычно одного из них в процентном соотношении гораздо больше.

Благодаря совместным усилиям биологов и программистов множество геномов уже расшифровано, поэтому есть такое понятие как "сборка по референсу"[19] – это способ ассемблирования, когда известен близкий к искомому геном и можно определить, какие риды из какого участка простым прикладыванием к известной последовательности. Но чаще либо нет такой близкой последовательности, либо нет информации об исследуемом организме. Сборка, полученная без опоры на известные геномы, называется сборкой de novo.

Ген – 1) невидимый, содержащий информацию элемент, эти элементы равномерно распределяются между двумя дочерними клетками при каждом клеточном делении[16]; 2) последовательность символов из алфавита {A, C, G, T}, обладающая специфичными свойствами, которые при этом не определяют её однозначно и варьируются в зависимости от особенностей организма и закономерностей, наблюдающихся в геномной последовательности. Каждый ген является участком генома, то есть его подстрокой.

Значимой или кодирующей в гене может быть как вся его последовательность, так и набор подстрок. Такая структура определяется строением клетки исследумеого организма. Все организмы делятся на две группы: эукариоты и прокариоты. Эукариоты – это организмы, в клетках которых есть ядро. К таким организмам относятся, например, люди, животные и растения. Прокариоты – это организмы, в клетках которых ядра нет,

например, бактерии, такие как кишечная палочка. Ген прокариотов представлен в виде строки, которая в то же время является кодирующей частью, тогда как ген эукариотов разделён на кодирующие и некодирующие участки, которые в нём чередются. Из кодирующих участков гена в дальнейшем получаются белки, которые являются основной функциональной частью организма на клеточном уровне.

1.2. Формулировка задачи и её особенности

Теперь, когда основные определения даны, можно сказать, какова была первоначальная формулировка задачи. Задача определяется как “поиск генов в метагеномных сборках de novo”. Как уже отмечалось ранее, “поиск” – это не совсем уместное слово, так как по факту задача состоит в построении предположений о том, какие последовательности из сборки могут являться генами. Чтобы строить эти предположения, надо предоставить метод, который даёт наиболее точный ответ, а значит, нужно придумать, как сравнивать различные методы решения поставленной задачи.

Основной особенностью является то, что в качестве входных данных берутся метагеномные сборки. Это значит, что нужно искать гены в разных организмах. Дополнительно стоит обратить внимание на слова “сборка” и “de novo”, которые означают, что искать необходимо в неупорядоченном наборе подстрок генома, при этом нет возможности сопоставить анализируемую сборку с уже существующей последовательностью. Эксперименты и исследования договорились проводить на сборках прокариотических организмов.

Таким образом сформулированная задача в данной работе делится на две хорошо разделимые подзадачи:

– сравнение методов поиска генов,

– поиск генов в метагеномных сборках de novo.

2. Сравнение методов поиска генов

2.1. Мотивация

Существуют решения задачи поиска генов с другими начальными данными: по референсу, по multi-cell сборке. Все они основываются на похожих идеях, время их работы насчитывает часы на вычислительных машинах большой мощности, поэтому многократный запуск с целью тестирования затруднён.

Все известные инструменты поиска генов выдают информацию, достаточную для того, чтобы получить FASTA файл с последовательностями предсказанных генов. FASTA – специальный формат для записи нуклеотидных последовательностей[23]. Остаётся неизвестным, как понять, что он предсказан верно, не прибегая к биологическому эксперименту, и какое предсказание можно назвать верным. Всё было бы намного проще, если бы в геномах не было мутаций, а для поиска генов был хотя бы один точный, пусть и долго работающий, алгоритм.

2.2. Общая идея инструемнта для сравнения

Решение – инструмент, который умеет устанавливать и запускать заданные методы поиска генов, преобразовывать результат их работы в FASTA и выравнивать[24] полученные последовательности на реально существующие гены. Тестирование необходимо проводить на сборках тех организмов, для которых истинные гены уже найдены. Таким образом, на вход инструменту необходимо подать FASTA файлы со сборками, реальные гены, известные для исследуемых организмов, реализации методов предсказывания генов с инструкциями запуска и установки и описанием выходных данных. По завершению работы этого инструмента хочется узнать, какие гены предсказаны верно, какие гены предсказаны неверно, а какие гены предсказать не удалось каждым из используемых методов, а также визуализировать эти результаты.

Если подойти к разработке этого инструмента именно так, то его можно использовать не только для методов предсказывания генов, но и для методов решения любых других задач, в результате решения которых можно получить FASTA файлы, и если для сравнения можно использовать достоверные данные.

2.3. Метрики

Для оценки качества поиска генов в этой работе будут использоваться понятия TP (True Positives), FP (False Positives), TN (True Negatives), FN (False Negatives), P (Precision) и R (Recall) для двух гипотез:

H_1 – предсказанный ген является реальным

H_2 – реальный ген предсказан

Гипотезы строятся для результатов работы каждого отдельного метода.

P и R считаются по следующим формулам

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

Подсчёт остальных параметров для каждой гипотезы ведётся по-своему.

Для гипотезы H_1:

TP – количество предсказанных генов, для которых нашлись похожие гены среди истинных данных, похожесь определяется границей точности используемого способа выравнивания;

FP – количество предсказанных генов, которые не нашлись в наборе реальных или нашлись, но выравнивание произошло с маленькой точностью;

TN – количество генов, которых не было в наборе реальных генов и не оказалось в числе предсказанных;

FN – количество генов, которые были среди реальных, но предсказать их не удалось.

Для гипотезы H_2:

TP – количество реальных генов, которые были предсказаны;

FP – количество реальных генов, которые предсказаны не были;

TN – количество генов, которых не было и которые при этом не были предсказаны;

FN – количество генов, которые были предсказаны, но среди реальных их не было.

Всё это необходимо считать отдельно, так как для каждой из гипотез будет использоваться свой способ выравнивания. Так можно добиться более точной оценки сравнения.

TN в обоих случаях явно посчитать невозможно, но нетрудно понять, что в обоих случаях это число равно бесконечности, так как генов, которых не было бесконечно много, так же как и генов, которые не удалось предсказать.

2.4. Реализация

Для простоты и наглядности, а также для возможного дальнейшего использования биологами инструмент реализован на языке Python с визуализацией на языке R. Для Python есть развивающаяся библиотека BioPython[5], которая умеет извлекать данные из FASTA файлов, а также производить многие другие операции над геномными последовательностями. Для языка R написана библиотека для простой визуализации больших объёмов данных ggpolt2[27], что также помогло наглядно представить полученные результаты сравнения различных методов.

Встраивание методов происходит в классовой структуре: есть базовый абстрактный класс для поиска генов, добавить метод можно, описав его поведение на языке методов этого базового класса.

Степень схожести предсказанных генов с реальными определяется выравниванием с использованием инструмента BWA[13]. Этот инструмент пытается приложить каждый элемент из одного набора ко всем элементам другого набора так, чтобы свести к минимуму операции вставки, удаления и замены при этом прикладывании. Так для каждого элемента из первого набора появляется один наиболее близкий элемент из второго набора. Несмотря на то, что выбирается наиболее близкий элемент, качество его приложения к первому может быть достаточно низким, тогда такое выравнивание не учитывается.

BWA-tool на выходе предоставляет SAM[21] файл, который показывает, какие элементы оказались схожими и с каким качеством. Если качество выравнивания (mapq) оказывается меньше 200, то такое выравнивание будем считать неудачным и просто говорить, что соответствие не нашлось.

После получения значений ранее описанных метрик для всех пар {инструмент, сборка}, строятся графики, описывающие описывающие разницу в количестве найденных и потерянных генов разными методами.

3. Поиск генов и метагеномные сборки

3.1. Общий подход и существующие решения

Существуют различные формулировки задачи поиска генов:

– в строке генома (в полной сборке)[17];

– в multi-cell сборке[10];

– в single-cell сборке[22];

– в метагеномной сборке[9].

Есть методы для решения поставленной задачи для первого пункта, которые используются также для решения задачи во второй формулировке. Эти методы подробнее будут описаны далее. Подход в данном случае заключается в том, что можно переиспользовать существующие решения для более сложных задач, тем более, что все эти методы являются лишь приближением.

Многие существующие решения – это некоторые обученные алгоритмы, которые при обучении ориентировались на определённые параметры рассматриваемого генома. Чаще всего для этого используется GC-состав или неявно указанная информация об организме и его геномной последовательности. GC-состав вычисляется как процент, который составляют гуанин (G) и цитозин (C) от длины исследумеой нуклеотидной последовательности.

Если вернуться к первоначально сформулированной задаче, цель данной работы – исследование задачи поиска генов именно в метагеномных сборках, то есть необходимо сформулировать гипотезы для решения этой задачи. Часто вместе с понятием о метагеномных сборках вспоминают о кластеризации, то есть прежде, чем данные как-то анализировать, предлагается их кластеризовать, то есть разбить контиги по организмам, но вспомним, что нам не известно, какие организмы находятся в сборке и сколько их там.

Как следствие всего, что описано выше, можно предполложить, что решение задачи будет происходить по одной из следующих двух схем:

1) кластеризовать контиги и воспользоваться одним из методов для решения задачи поиска генов для референсной последовательности;

2) поиск генов без предварительной кластеризации.

Теперь, прежде чем сравнивать эти два подхода, необходимо разобраться с кластеризацией контигов в метагеномной сборке и выяснить, как уже существующие методы генов работают в каждой из предложенных ситуаций без предварительной обработки данных, то есть без выявления особенностей конкретного генома. В реализации второго пункта используется инструмент для сравнения поиска генов, описанный ранее.

3.2. Описание предлагаемой кластеризации контигов в метагеномной сборке

3.2.1. Метод k-средних

Суть метода k-средних[26] заключается в поиске множества C центров кластеров (центроидов) $c \in \mathbb{R}^m$, где $|C| = k$, такого, что для всего набора данных X, $x \in \mathbb{R}^m$, решается задача о минимизации следующей функции

$$\min \sum_{x \in X} \|d(C, x) - x\|^2 \tag{3}$$

Здесь $d(C, x)$ считается как евклидово расстояние от x до ближайшего к нему центра $c \in C$.

Алгоритм.

1. Произвольно выбираются k центроидов $C = \{c_1, \ldots, c_k\}$.

2. Для каждого $i \in \{1, \ldots, k\}$, создаётся кластер C_i, который состоит из множества точек из X, которые ближе к c_i, чем ко всем c_j, таким что $j \neq i$.

3. Для кадого $i \in \{1, \ldots, k\}$, находим в качестве центра такой c_i, что он является центром масс для множества точек в C_i:

$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{4}$$

4. Повторять шаги 2 и 3 до тех пор, пока не перестанет меняться.

3.2.2. Метод k-средних++

Результаты кластеризации по методу k-средних зависят от выбора начальной конфигурации центроидов, то есть от инициализации, работа алгоритма существенно замедляется при кластеризации больших объёмов данных, алгоритм может сходиться к локальному минимуму целевой функции.

Чтобы избавиться от этих проблем, в алгоритме k-сердних++[2] усовершенствован метод инициализации центров. $D(x)$ далее обозначает кратчайшее расстояние от точки x до ближайшего выбранного центра.

a. С нормальной вероятностью выбирается центр c_1 из X.

b. Выбирается следующий центр $c_i = x' \in X$ с вероятностью

$$\frac{D(x')^2}{\sum\limits_{x \in X} D(x)^2} \tag{5}$$

c. Шаг b повторяется, пока центров не станет k.

Метод k-средних++ меняет только инициализацию, и шаги 2-4 повторяются.

3.2.3. Быстрый метод k-средних

Так как объёмы данных в поставленной задаче большие, а для экспериментов важно уметь их быстро воспроизводить, для решения поставленной задачи был использован не только метод k-средних++, но и версия алгоритма, называющаяся быстрым методом k-средних (Fast k-means или Mini-batch k-means[20])

3.2.4. Подбор k

В задаче кластеризации количество кластеров заранее неизвестно, поэтому k необходимо подобрать. Подбор k осуществляется, максимизурая байесовский информационный критерий[6].

3.2.5. Алгоритм кластеризации

Параметр, по которому производится кластеризация, выбран эмперически. Этим параметром стали частоты встречаемости 6-меров на длину всего генома. n-мер – это последовательность нуклеотидов длины n, в данном случае $n = 6$. Такой параметр оказался лучше в сочетании точности и скорости работы, чем кластеризация по GC-содержанию и матрицы перехода.

Алгоритм кластеризации

1. Подсчитываются частоты встречаемости в геноме всех 6-мер.
2. Применяется метод главных компонент.
3. Запускается быстрый алгоритм k-средних с выбором k по байесовскому информационному критерию.

3.3. Существующие методы поиска генов и их улучшения

3.3.1. Поиск открытых рамок считывания

Поиск открытых рамок считывания – это самый простой, долго работающий и неточный метод определения генов, но для полной последовательности генома он должен покрыть все гены, хоть и с большим излишком. Открытая рамка считывания или ORF (open reading frame) – последовательность нуклеотидов, потенциально способная кодировать белок, то есть, возможно, являющаяся геном. С точки зрения самой последовательности это строка, которая начинается со старт-кодона и закачивающаяся стоп кодоном. Старт-кодоны и стоп-кодоны – это определённые трёхбуквенные последовательности, обозначающие начало кодирования возможного белка и конец соответственно. Иногда считается, что стоп-кодонов может и не быть – тогда слово "открытая" в названии полностью себя оправдывает. При поиске октрытых рамок считывания учитывается, что эти последовательности могут начинаться с любого места генома, а также что стоп-кодон до следующего начала может быть не один.

3.3.2. GeneMark

GeneMark[14] – самый первый инструмент для предсказывания генов из группы GeneMark. Основная идея метогда, заложенного внутри этого инструмента, – сочетание марковской модели для определения кодирующих и некодирующих участков с Байесовской функцией принятия решений. Метод можно обобщить и на марковские цепи высших порядков. Далее эта идея расширяется тем, что поиск генов осуществляется по двум стрендам. Стренд – одна из двух цепочек ДНК, они друг другу комплиментарны и определяются по тому правилу, что аденин (A) соединяется комплементарной связью с тимином (T), а гуанин (G) соединяется комплементарной связью с цитозином (C). Проблема с поиском генов на компелементарном стренде заключается в том, что многие методы находят гены на обеих цепочках одновременно, а так быть не должно. Здесь для того, чтобы избежать такого явления, гены ищутся на прямом стренде, затем участки, комплементарные тем, на которых нашлись гены, называются затенёнными и в поиске не участвуют.

Выбор модели, по которой производится поиск генов осуществляется на основании подсчёта GC-содержания рассматриваемых данных. При использовании инструмента выбирается соответствующая матрица. Матрицу можно сгенрировать на основе каких-то своих данных или использовать набор матриц, предлагаемых авторами.

Видно, что метод основан на том, что есть полная геномная последовательность. Как быть в случае сборки? Можно пойти по одному из двух путей: считать GC-содержание по всем контигам вместе и по каждому контигу в отдельности. Так как разница эффективности этих двух подходов нигде не описана, при проведении эксперимента проверяются обе версии.

3.3.3. GeneMark.HMM

GeneMark.hmm[1] – это улучшенная версия GeneMark, которая тоже опирается в выборе модели на GC-содержание рассматриваемого организма, поэтому для него тоже будут рассмотрены два варианта.

Алгоритм GeneMark.hmm разработан для поиска более точных границ генов. Идея заключается в встраивании в метод обыкновенного GeneMark поиск чётких границ генов.

Рамки генов моделируются как переходы между скрытыми состояниями. Эти модели, как и в предыдущем случае, можно сгенерировать самостоятельно на основании известных пользователю данных, а можно взять предложенные авторами. В этом методе также добавлено отсечение по местам инициализации кодирования последующего участка, однако в случае сборки это можно быть как полезно, так может и негативно отразиться на точности.

3.3.4. GeneMark-S

GeneMark-S[4] – это надстройка над методом GeneMark.hmm, явно использующая его в реализации. Этот метод основан на обучении без учителя и работает без предварительных знаний о каких-либо протеин-кодирующих участках генома. Здесь совмещены модели для кодирующих и некодирующих регионов и модели для поиска участков генома, находящихся незадолго до последовательности генов и биологически определяющих начало процесса кодирования.

На полногеномных последовательностях этот метод является лучшим из всех методов группы GeneMark.

3.4. Эксперимент

На данном этапе исследования было принято решение проверить две гипотезы:

- предложенный метод кластеризации можно использовать для метагеномных сборок,
- можно выявить лучший инструмент на основе HMM для дальнейшего использования в исследовании метагеномных сборок.

3.4.1. Суть эксперимента

Запуск и сравнение существующих методов поиска генов без подготовки в случаях полной последовательности, multi-cell сборки, single-cell сборки. Сравнение и визуализация результатов осуществляется с помощью разработанного и описанного выше инструмента.

Поиск оптимальных параметров для кластеризации, который заключается в тестировании кластеризации на данных, описанных далее с использованием матриц переходов и частот встречаемости k-мер.

Кластеризация описанным методом всех рассматриваемых сборок. Ожидается, что для не метагеномных сборок в результате будет ярко выражен один крупный кластер, результаты визуализируются гистограммой по размерам кластеров.

3.4.2. Тестовые данные

Все ниже описанные данные рассмотрены в сочетании с их полными последовательностями и наборами реальных генов.

Полные последовательности геномов. В качестве тестовых данных для запуска поисковиков генов взяты последовательности организмов, описанных в таблице 1

Таблица 1: Полные геномы, используемые в данной работе

Name	GC%	Gene Count
Escherichia coli K12 substr. MG1655[11]	50.79	4466
Pedobacter heparinus[8]	42.05	4287
Meiothermus ruber[7]	63.38	3105

Метагеномные данные. В качестве тестовых данных этого рода взята искуственно сгенерированная сборка[3], содержащая организмы, описанные в таблице 2.

Single-cell[12] сборки. Эти данные получены при помощи ассемблера SPAdes. Сборки представляют организмы, описанные в таблице 1. Всего расстмотрено одиннадцать single-cell сборок.

Multi-cell сборки. Для тестирования инструментов на сборках такого вида использовались данные организмов, описанных в таблице 3. Сборка для Escherichia coli получена ассемблером SPAdes.

3.5. Результаты эксперимента

3.5.1. Сравнение результатов поиска генов

Рис. 1, 2, 3, 4, 5, 6, 7, 8, 9.

3.5.2. Кластеризация

Рис. 10, 11, 12, 13.

3.5.3. Пояснение результатов

В результате для сравнения методов предсказывания генов стоит обращать внимание как на то, какие гены были предсказаны, а какие нет, так и на то, какие из предсказанных генов существуют.

Так, например, графики для Precision/Recall различают две гипотезы: предсказанный ген существует, существующий ген предсказан. Если рассматривать эти две гипотезы в совокупности, то GeneMark-S и GeneMark.hmm EGC (с вычислением GC-содержания для каждой последовательности отдельно) решают задачу поиска генов лучше других методов, но при этом GeneMark.hmm находит меньше несуществующих генов.

Метод кластеризации получился неточным, однако есть предположение, что это можно исправить, выбрав другую метрику: использовать метрику по Interpolated Markov Model[15], что позволит учитывать особенности даже в коротких нуклеотидных последовательностях.

Заключение

В результате данной работы придумана и разработана основа инструмента для сравнения методов анализа геномных последовательностей. На этой основе реализован инструмент для сравнения методов поиска генов.

Предложено улучшение методов группы GeneMark, опирающихся в своих моделях на GC-содержание нуклеотидных последовательностей, показана эффективность этого улучшения. Улучшенный метод GeneMark.hmm с подсчётом GC-содержания для каждого контига в отдельности реализован в проекте QUAST[18].

Предложен метод кластеризации последовательностей из метагеномной сборки, проанализирована его точность на сборках различного типа, предложена гипотеза, что кластеризацию можно улучшить, выбрав в качестве параметра Interpolated Markov Model.

Показано, что среди методов поиска генов, основанных на HMM, GeneMark.hmm с модификацией работает лучше остальных методов на прокариотических геномах, на single-cell сборках и на multi-cell сборках.

Список литературы

[1] A. Lukashin, M. Borodovsky. GeneMark.hmm: new solutions for gene finding. // Nucleic Acids Research. — 1998. — Vol. 26, no. 4. — P. 1107–1115.

[2] Arthur David, Vassilvitskii Sergei. k-means++: the advantages of careful seeding // Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. — SODA '07. — Philadelphia, PA, USA : Society for Industrial and Applied Mathematics, 2007. — P. 1027–1035.

[3] Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data / Daniel R. Mende, Alison S. Waller, Shinichi Sunagawa et al. // PLoS ONE. — 2012. — 02. — Vol. 7, no. 2. — P. e31386.

[4] Besemer J. Lomsadze A., M. Borodovsky. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. // Nucleic Acids Research. — 2001. — Vol. 29, no. 12. — P. 2607–2618.

[5] BioPython. — 2013. — URL: `http://biopython.org` (online; accessed: 25.05.2013).

[6] Chen Scott, Gopalakrishnan Ponani. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion // Proc. DARPA Broadcast News Transcription and Understanding Workshop / Virginia, USA. — 1998. — P. 8.

[7] Complete genome sequence of Meiothermus ruber type strain (21T) / Brian J Tindall, Johannes Sikorski, Susan Lucas et al. // Standards in genomic sciences. — 2010. — Vol. 3, no. 1. — P. 26.

[8] Complete genome sequence of Pedobacter heparinus type strain (HIM 762-3T) / Cliff Han, Stefan Spring, Alla Lapidus et al. // Standards in Genomic Sciences. — 2009. — Vol. 1, no. 1. — P. 54.

[9] De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities / Priya Narasingarao, Sheila Podell, Juan A Ugalde et al. // The ISME journal. — 2011. — Vol. 6, no. 1. — P. 81–93.

[10] Design of a three-dimensionally controlled multi-cell-assembly system based on the control of a mixer nozzle. / Q Wang, M Xu, Y Li et al. // Journal of biomedical engineering. — 2011. — Vol. 28, no. 5. — P. 1030.

[11] Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. / Riley M, Abe T, Arnaud MB et al. // Nucleic acids research. — 2006.

[12] Lasken Roger S. Single-cell genomic sequencing using multiple displacement amplification // Current opinion in microbiology. — 2007. — Vol. 10, no. 5. — P. 510–516.

[13] Li H., Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform // Bioinformatics. — 2010. — Mar. — Vol. 26, no. 5. — P. 589–595.

[14] M. Borodovsky, J. McIninch. GeneMark: parallel gene recognition for both DNA strands. // Computers & Chemistry. — 1993. — Vol. 17, no. 19. — P. 123–133.

[15] Microbial gene identification using interpolated Markov models / Steven L Salzberg, Arthur L Delcher, Simon Kasif, Owen White // Nucleic acids research. — 1998. — Vol. 26, no. 2. — P. 544–548.

[16] Molecular Biology of the Cell, 4th edition / Bruce Alberts, Alexander Johnson, Julian Lewis et al. — New York: Garland Science, 2002.

[17] Mount David W. Sequence and genome analysis // Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour. — 2004. — Vol. 2.

[18] QUAST: quality assessment tool for genome assemblies / Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, Glenn Tesler // Bioinformatics. — 2013. — Vol. 29, no. 8. — P. 1072–1075. — URL: http://bioinformatics.oxfordjournals.org/content/29/8/1072.abstract.

[19] Reference-assisted chromosome assembly / Jaebum Kim, Denis M. Larkin, Qingle Cai et al. // Proceedings of the National Academy of Sciences. — 2013. — URL: http://www.pnas.org/content/early/2013/01/09/1220349110.abstract.

[20] Sculley D. Web-scale k-means clustering // Proceedings of the 19th international conference on World wide web. — WWW '10. — New York, NY, USA : ACM, 2010. — P. 1177–1178.

[21] The Sequence Alignment/Map format and SAMtools / Heng Li, Bob Handsaker, Alec Wysoker et al. // Bioinformatics. — 2009. — Vol. 25, no. 16. — P. 2078–2079.

[22] Spades: A new genome assembly algorithm and its applications to single-cell sequencing / Anton Bankevich, Sergey Nurk, Dmitry Antipov et al. // Journal of Computational Biology. — 2012. — Vol. 19, no. 5. — P. 455–477.

[23] Wikipedia. FASTA format // Wikipedia, The Free Encyclopedia. — 2013. — URL: http://en.wikipedia.org/wiki/FASTA_format (online; accessed: 05.05.2013).

[24] Wikipedia. Sequence alignment // Wikipedia, The Free Encyclopedia. — 2013. — URL: http://en.wikipedia.org/wiki/Sequence_alignment (online; accessed: 23.05.2013).

[25] Wikipedia. Strain (biology) // Wikipedia, The Free Encyclopedia. — 2013. — URL: http://en.wikipedia.org/wiki/Strain_(biology) (online; accessed: 23.05.2013).

[26] Xu Rui, Wunsch Don. Clustering. — Wiley-IEEE Press, 2009. — ISBN: 9780470276808.

[27] ggplot2. — URL: http://ggplot2.org (online; accessed: 15.05.2013).

Таблица 2: Организмы, используемые для генерации метагеномной сборки

Name	phylo group	genome size (Mb)	GC %
Cyanothece sp. ATCC 51142	Cyanobacteria	5.43	37.9
Staphylococcus aureus subsp. aureus str. Newman	Firmicutes	2	32.9
Methanococcus maripaludis C7	Euryarchaeota	1.8	33.3
Neisseria meningitidis MC58	Betaproteobacteria	2.3	51.5
Bacillus clausii KSM-K16	Firmicutes	4.3	44.8
Escherichia coli str. K12 substr. W3110	Gammaproteobacteria	4.65	50.8
Listeria welshimeri serovar 6b str. SLCC5334	Firmicutes	2.8	36.4
Lawsonia intracellularis PHE/MN1-00	Deltaproteobacteria	1.76	33.1
Thiobacillus denitrificans ATCC 25259	Betaproteobacteria	2.91	66.1
Rhodopseudomonas palustris CGA009	Alphaproteobacteria	5.47	65

Таблица 3: Организмы, multi-cell сборки которых использованы для определения качества поиска генов на подобных данных

Name	GC%	Gene Count
Escherichia coli	50.79	4466

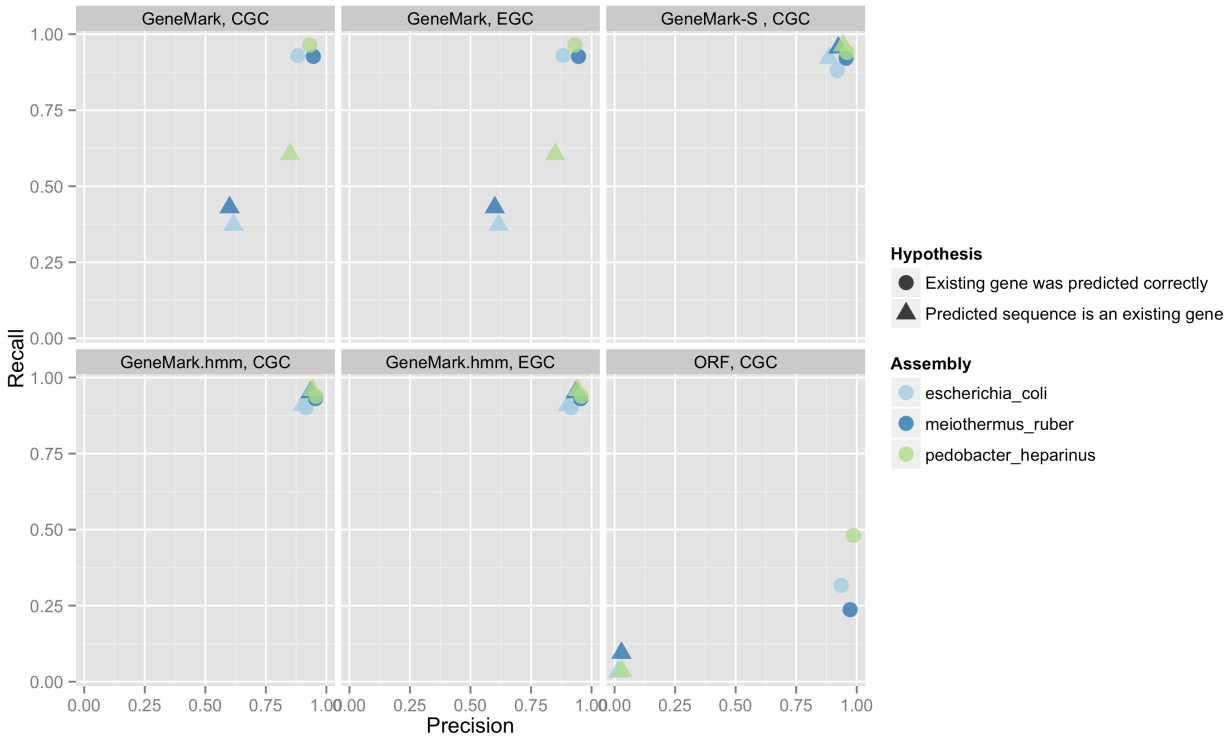

Рис. 1: P/R для полногеномных последовательностей

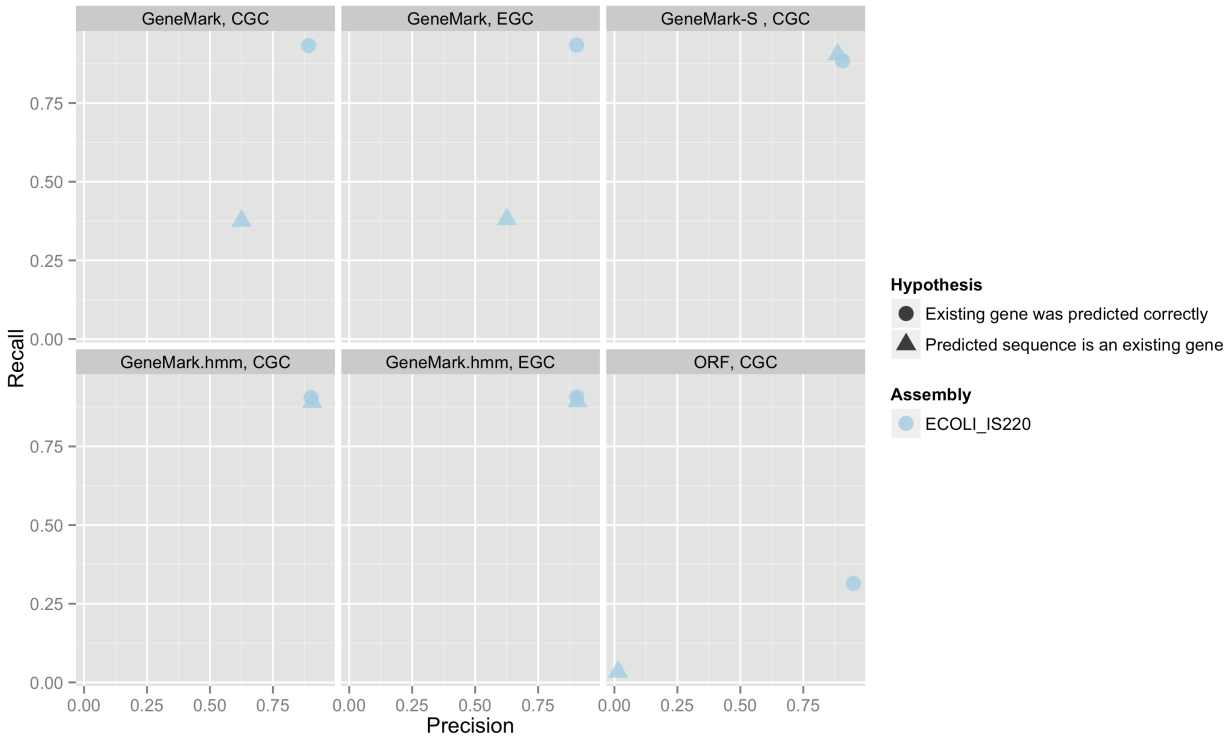

Рис. 2: P/R для multi-cell сборок

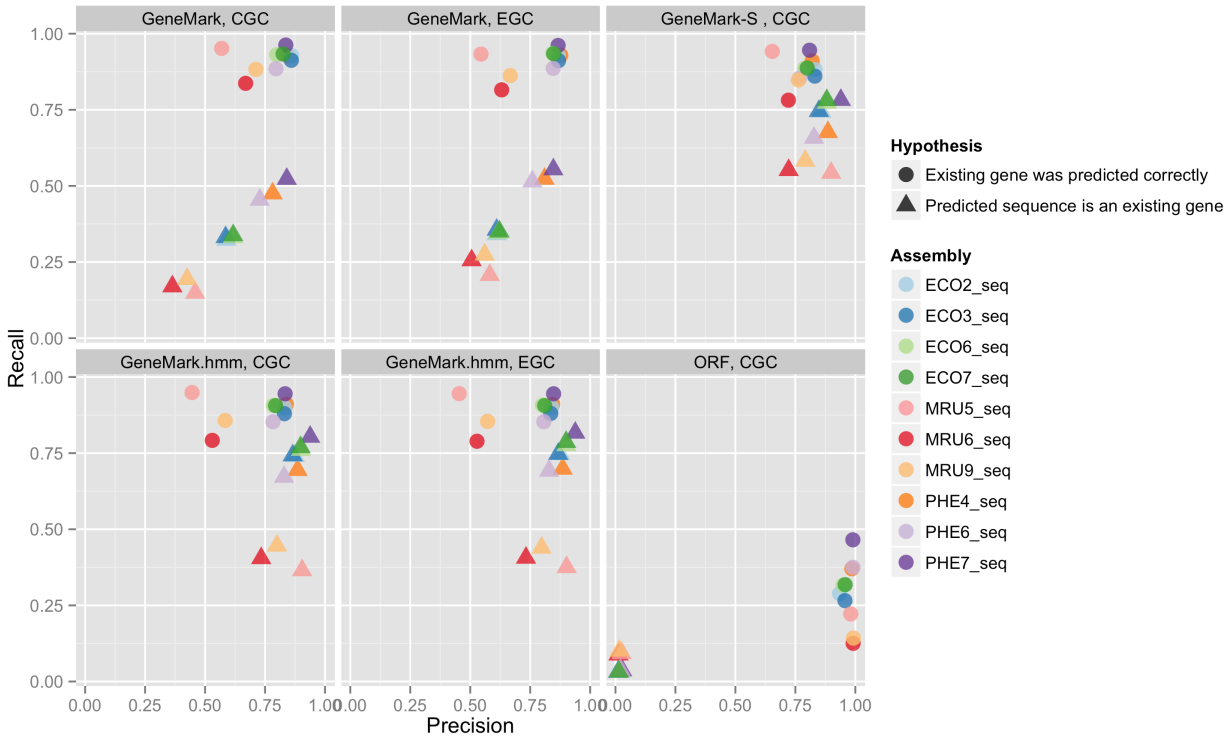

Рис. 3: P/R для single-cell сборок

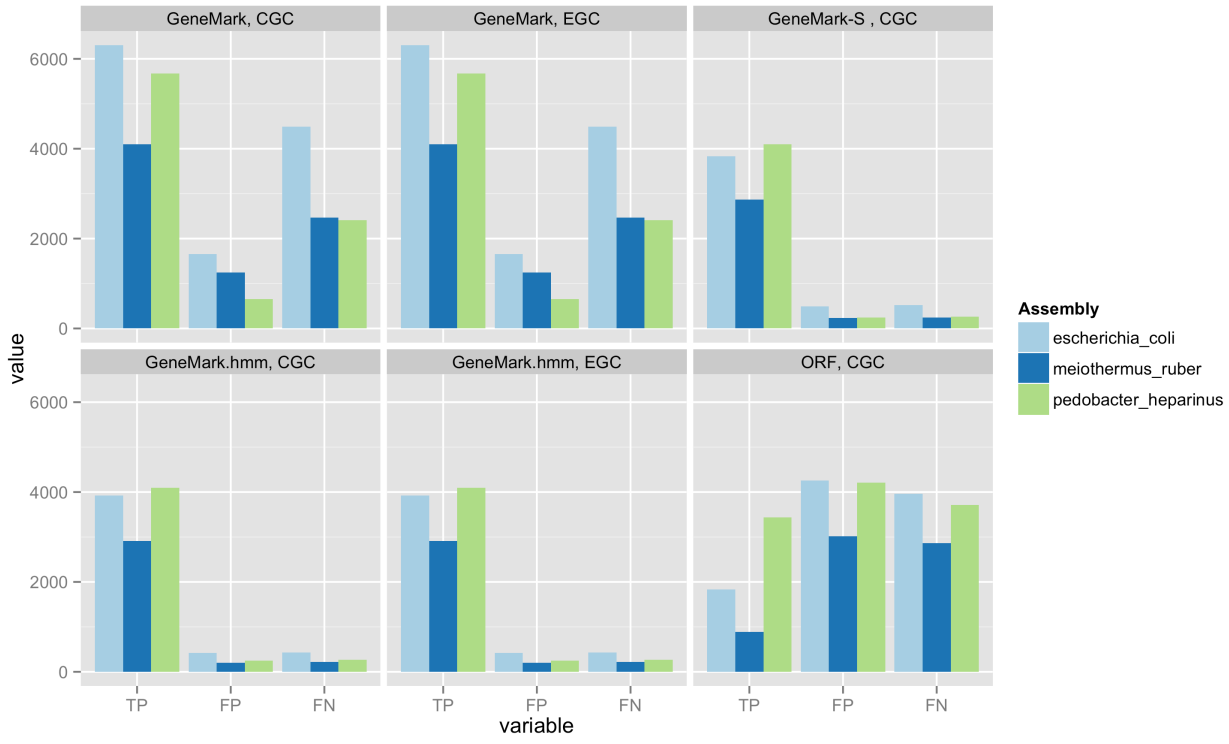

Рис. 4: TP, FP, FN для полногеномных последовательностей

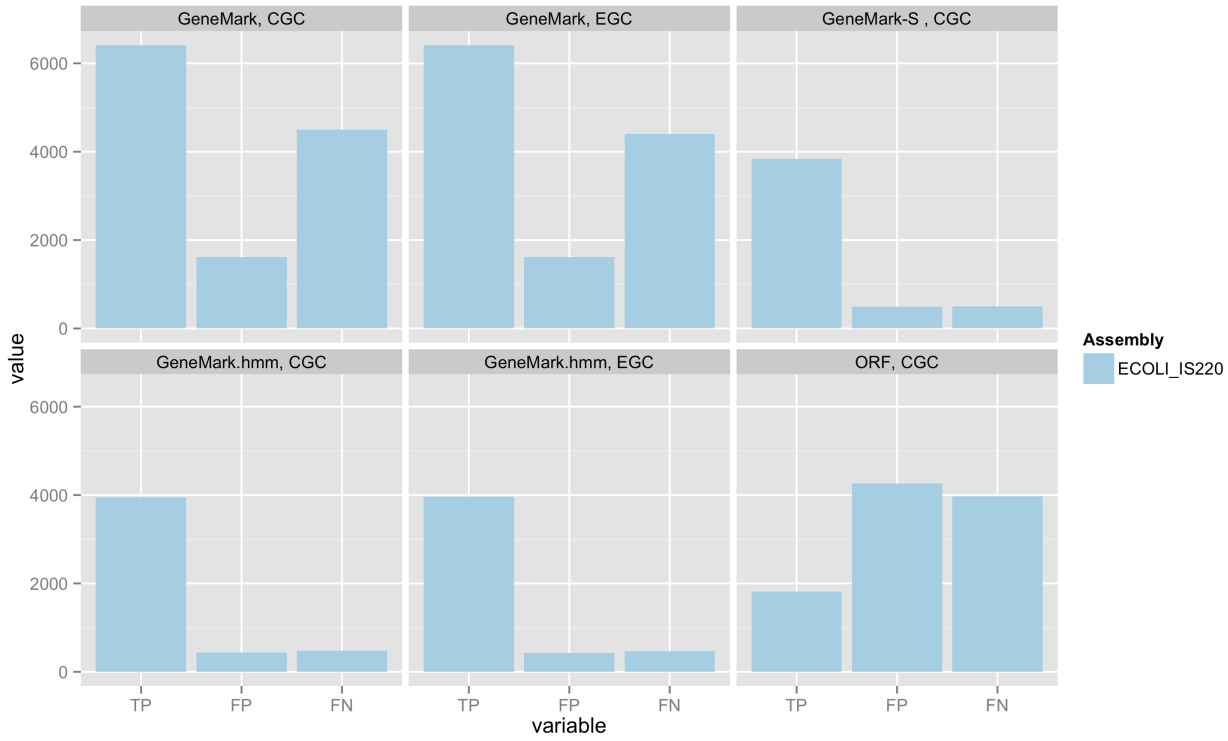

Рис. 5: TP, FP, FN для multi-cell сборок

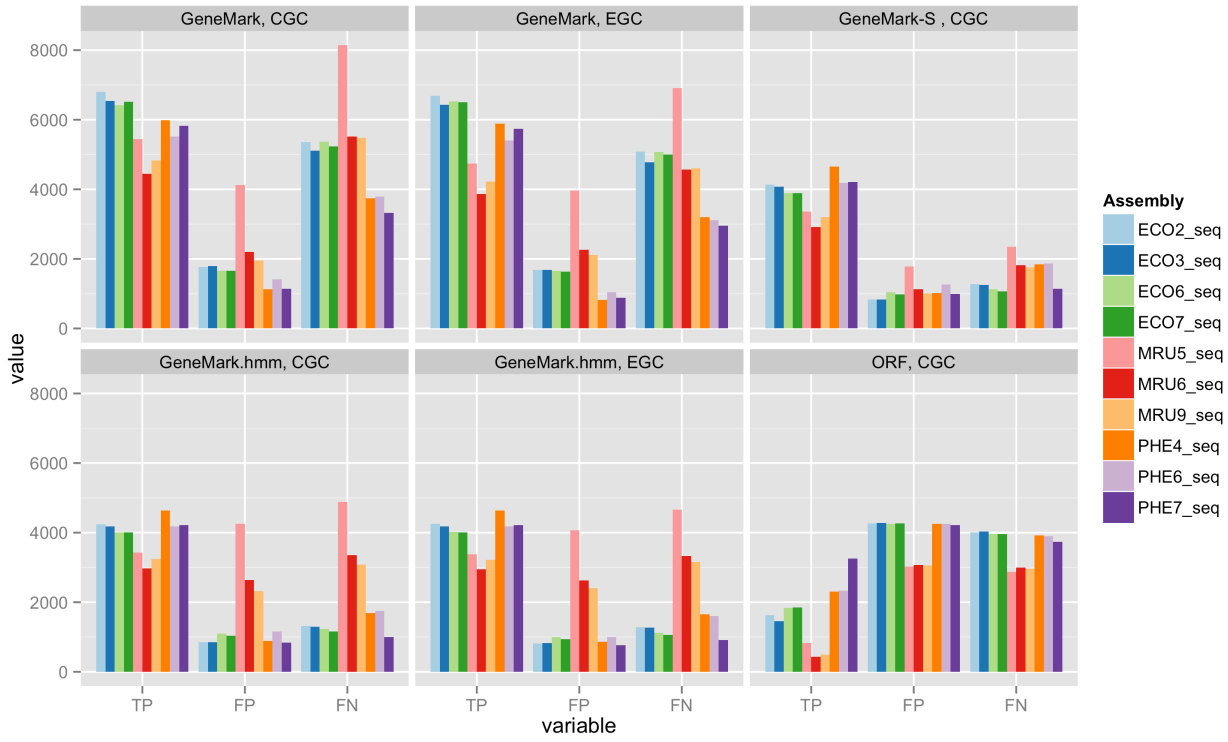

Рис. 6: TP, FP, FN для single-cell сборок

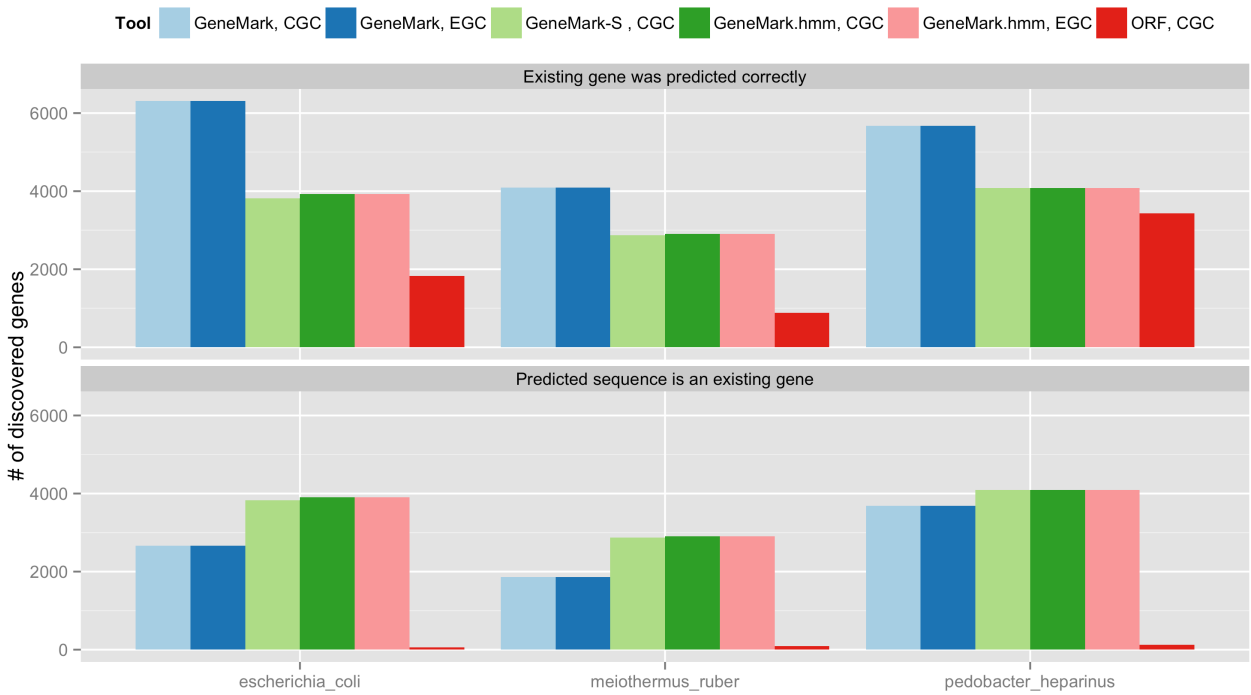

Рис. 7: Точность поиска генов для полногеномных последовательностей

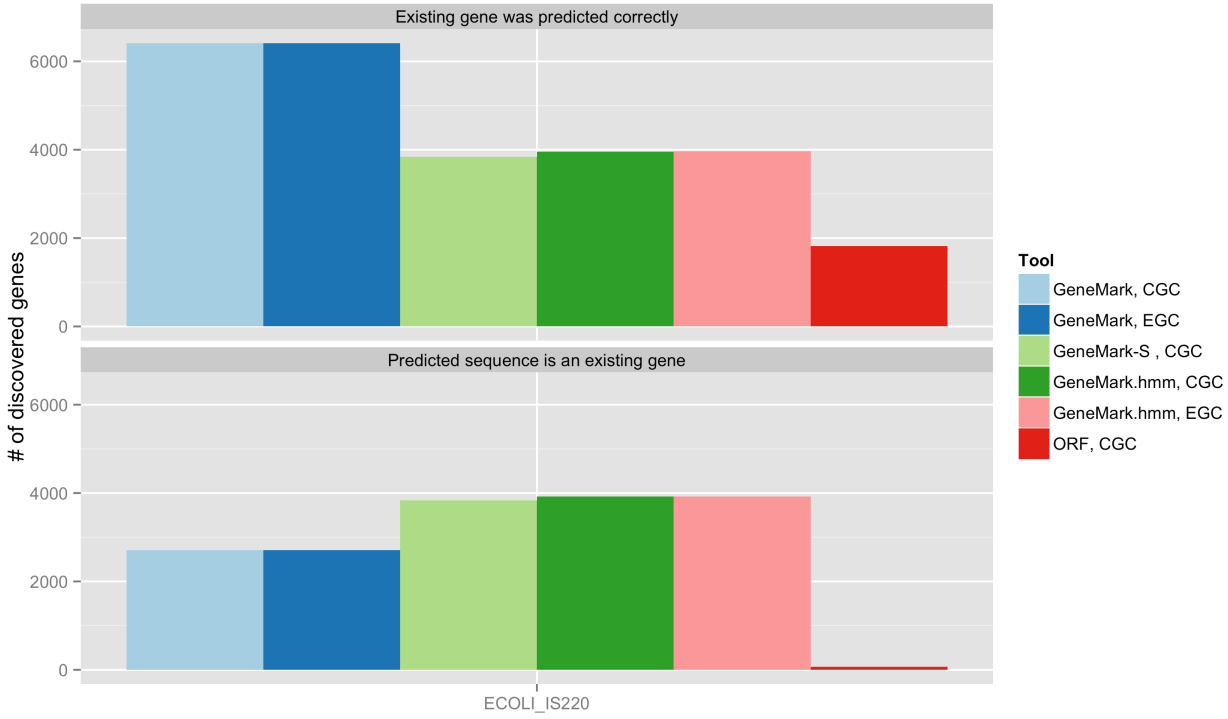

Рис. 8: Точность поиска генов для multi-cell сборок

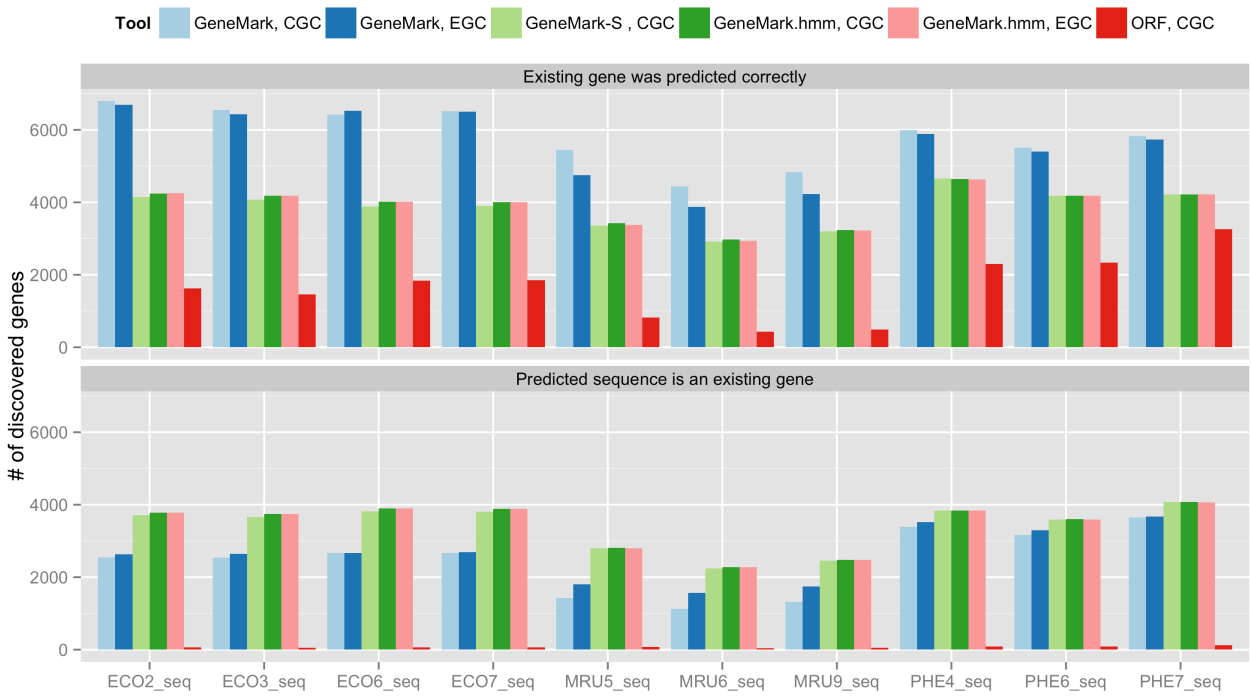

Рис. 9: Точность поиска генов для single-cell сборок

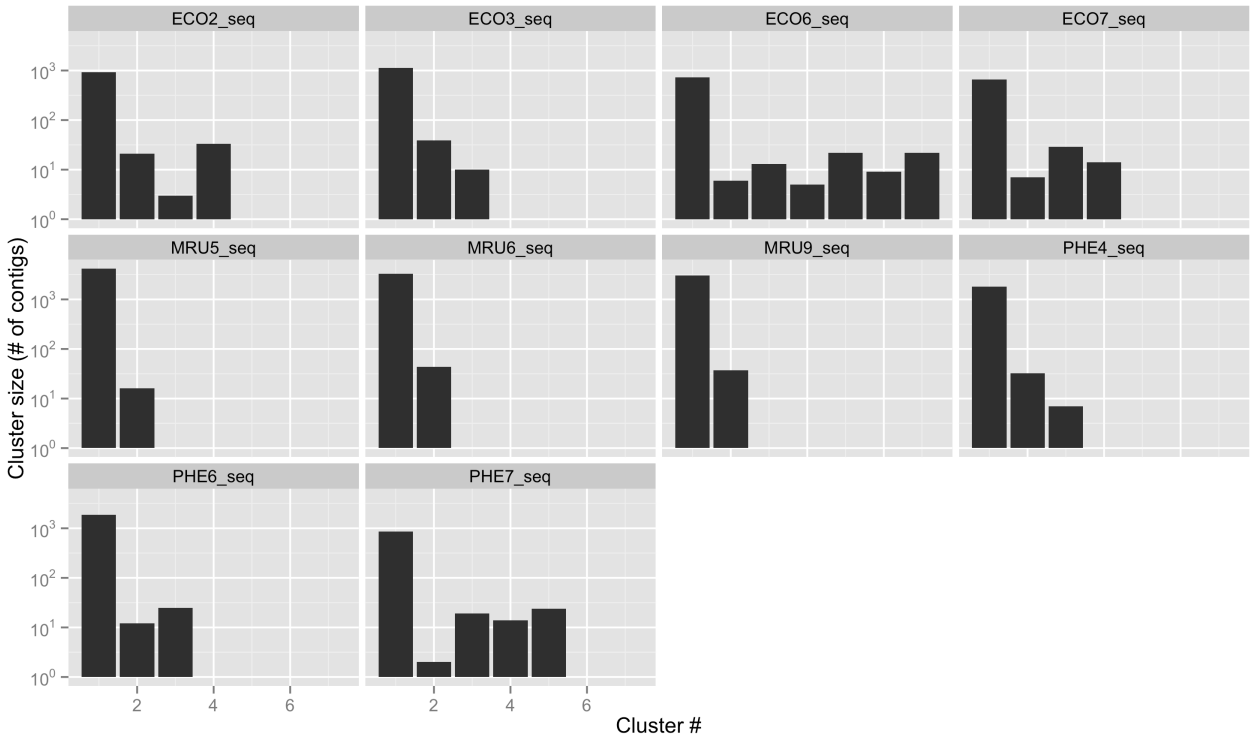

Рис. 10: Кластеризация последовательностей из single-cell сборок по 6-мерам

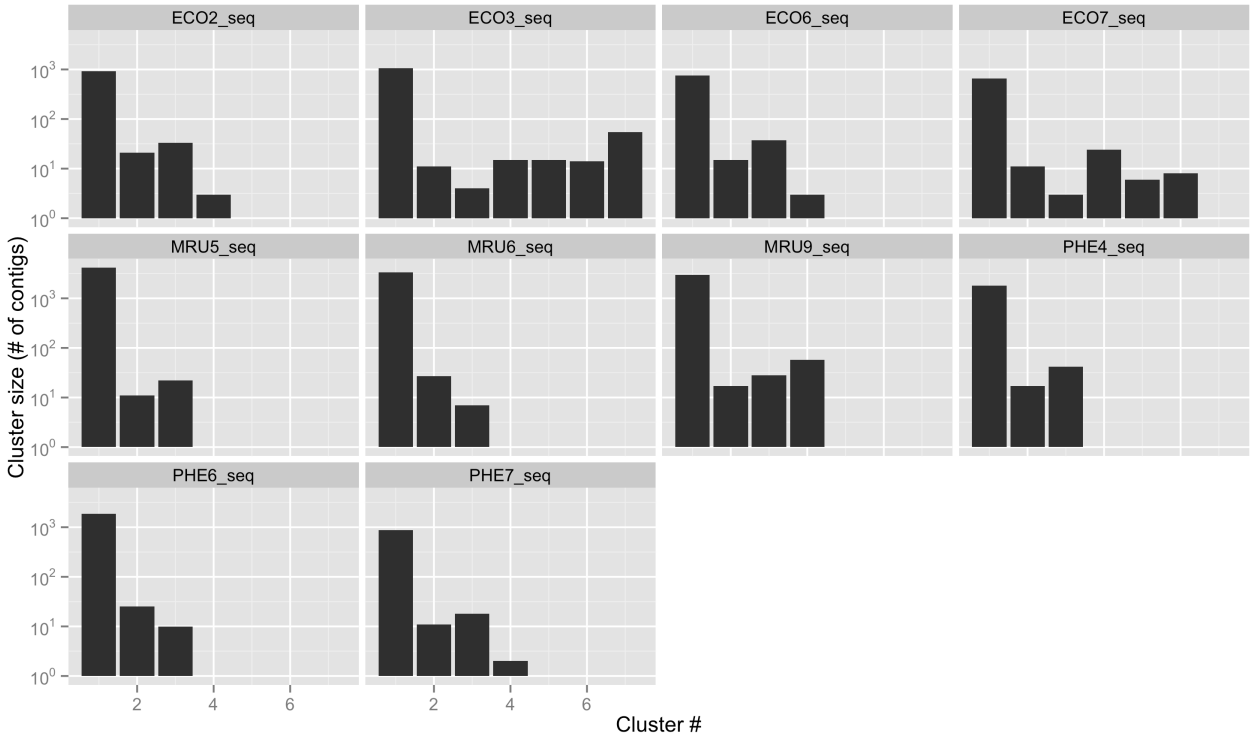

Рис. 11: Кластеризация последовательностей из single-cell сборок по 8-мерам

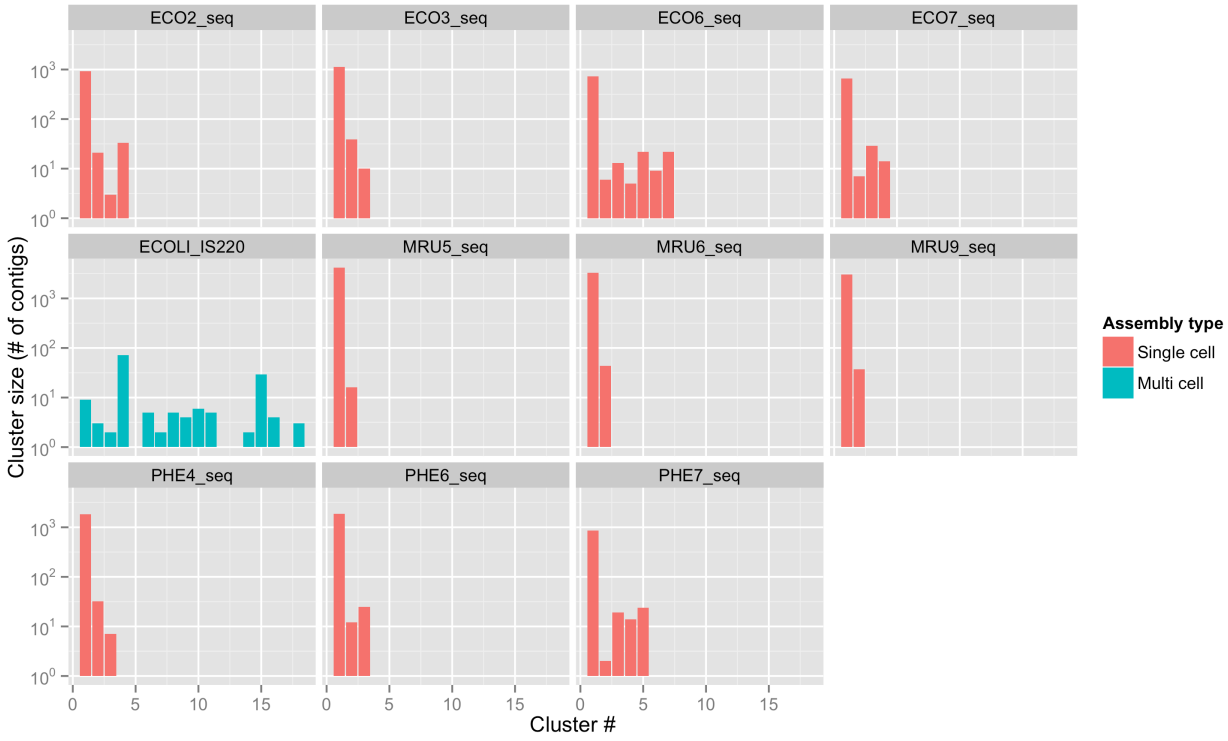

Рис. 12: Кластеризация последовательностей из single-cell сборок и multi-cell сборки по 8-мерам

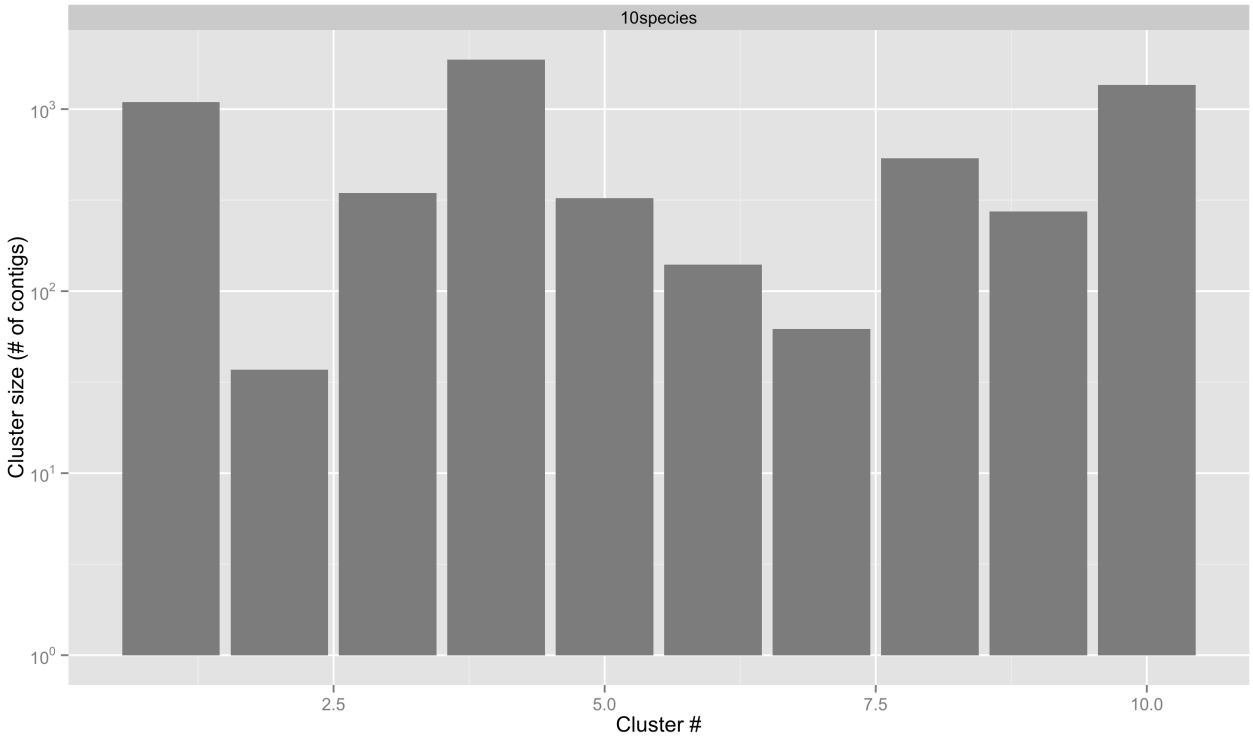

Рис. 13: Кластеризация последовательностей из метагеномной сборки по 8-мерам, на гистограмме показаны только верно определённые кластеры